Can Statistics Do without Artefacts?

Jean-Bernard Chatelain

Can Statistics Do without Artefacts?

Jean-Bernard Chatelain[1]

Prisme N° 19

December 2010

[1] Jean-Bernard Chatelain is Professor of Economics at the Sorbonne Economic Centre (Université Paris-I) and Co-director of the CNRS European Research Group on Money, Banking and Finance. His research interests are growth, investment, banking and finance, applied econometrics and statistical methodology.

Summary

This article presents a particular case of spurious regression, when a dependent variable has a coefficient of simple correlation close to zero with two other variables, which are, on the contrary, highly correlated with each other. In these spurious regressions, the parameters measuring the size of the effect on the dependent variable are very large. They can be "statistically significant". The tendency of scientific journals to favour the publication of statistically significant results is one reason why spurious regressions are so numerous, especially since it is easy to build them with variables that are lagged, squared or interacting with another variable. Such regressions can enhance the reputation of researchers by stimulating the appearance of strong effects between variables. These often surprising effects are not robust and often depend on a limited number of observations, fuelling scientific controversies. The resulting meta-analyses, based on statistical synthesis of the literature evaluating this effect between two variables, confirm the absence of any effect. This article provides an example of this phenomenon in the empirical literature, with the aim of evaluating the impact of development aid on economic growth.

Contents

Introduction

Linear regression is one of the most widely used statistical methods in the applied sciences. It is used to evaluate whether and to what extent an increase in one variable (for example, development aid) is associated with a positive or negative effect on another variable (for example, economic growth). Its origins can be traced back to the least squares method (Legendre, 1805). It was then associated with the correlation between two normal distributions by Francis Galton (1886) in the case of what is called "simple regression". The extension to partial correlations of a dependent variable with at least two other variables was invented by George Udny Yule (1897). It is called "multiple regression".

This text challenges the validity of certain results obtained by the method of linear regression. I begin by adding to an already long list (Aldrich, 1995) a new case of spurious regression, where the method of linear regression indicates a connection between several variables, when in fact this connection cannot be verified. This type of regression appears to be a very particular, even trivial, case, which researchers should only rarely encounter in their work. I argue that the method for selecting papers using linear regression presents an artefact favouring the publication of these spurious regressions.

To make my case, I will start by addressing the question of statistical inference: how, on the basis of a certain number of observations, can one derive a certain probability about the relations that can exist between several variables. Ronald Fisher (1925) proposed a method of inference adapted to the technique of regression, which has gradually become established among researchers. Although it is extremely widespread in most scientific communities, it has been the subject of many criticisms that I shall briefly recall.

In the third section, I describe the sort of research practices that have been generated by the adoption of Fisher's method of statistical inference. They have led the editors of scientific journals to only publish results in which statistical inference rejects the so-called null hypothesis of an absence of connection between two variables. This problem leads to an artefact of publication whereby "negative" results, indicating the absence of relations between variables, are not published.

The particular case of spurious regression mentioned here then becomes very interesting. It enables researchers to find parameters (measuring the statistical

connection between two variables) that (1) are very high, (2) reject the null hypothesis of the absence of connection between the variables, (3) have an estimated value and sign that are particularly sensitive to the addition or removal of a few observations (which stimulates controversy and boosts the researcher's renown) and (4) reveal a previously unsuspected connection between these variables, which is much appreciated in the most prestigious scientific journals.

In the fourth section, I suggest two simple remedies: the parameter inflation factor (PIF), which measures the size of the effect between two variables, and hypothesis testing on the coefficients of simple correlations. I apply these indicators to a study of the effects of development aid on economic growth that has been very widely cited over the last ten years. With the help of these two indicators, I shall prove that this study presents a spurious regression.

1. The neglected problem of spurious regressions

a. An artefact of multiple regression

Simple regressions estimate a linear relation between two variables observed N times: for example, the growth of gross domestic product (GDP) per capita and development aid as a proportion of GDP observed for N countries over a given period.

To present this problem, I shall consider "standardized" variables, which have a mean of zero and a standard deviation of one. Standard deviation is a measurement of the dispersion of observations around the mean. It is always possible to standardize variables by subtracting the mean from each observation and dividing by the standard deviation. For standardized variables, the method of simple regression estimates a linear relation between two variables. The parameter of this relation is the simple correlation coefficient, as defined by Galton (1886). The higher the absolute value of this correlation coefficient, the greater the "size of the effect" of one variable on the other. At most, the absolute value is equal to one. Here are three examples of simple regression:

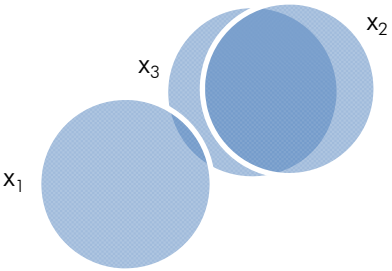
$$x_2 = 0.99\, x_3 + \varepsilon_{2.3}$$

$$x_1 = 0\, x_2 + \varepsilon_{1.2}$$
$$x_1 = 0.14107\, x_3 + \varepsilon_{1.3}$$

The variable on the left of the equation is the "dependent" or "explained" variable. The variable on the right of the equation is the "explanatory" variable. Measurement errors and omitted variables are taken into account through the "disturbances" denoted $\varepsilon_{i.j}$. Their numerical value for each observation in the sample is called the "residual".

The correlation coefficient can be interpreted as follows. When the variable x_3 deviates from its mean by one standard deviation, then the variable x_2 deviates from its own mean by 0.99 times its standard deviation. The two variables are highly correlated. For the second equation, on the other hand, when the variable x_2 deviates from its mean by one standard deviation, the variable x_1 does not deviate from its mean. The two variables are not correlated at all. Lastly, the variables x_1 and x_3 are very weakly correlated. These correlations can be represented in the form of a Venn diagram (Figure 1.1), where the circle representing x_2 largely overlaps with the circle representing x_3 without having any intersection with the circle representing x_1, which in turn has a small intersection with the circle x_3.

Figure 1.1: Venn diagram for the three simple correlations



Analysis of the variance of the dependent variable completes this information. The variance is the square of the standard deviation, which is denoted σ. In the case of a standardized variable, the variance is equal to one. Analysis of variance divides the dispersion of observations into different components according to whether it is predicted by the dispersion of the explanatory variable (variance explained by the

model) or by the related disturbance, for example, measurement errors or other unobserved phenomena (residual variance). The coefficient of determination is then defined as the ratio of the variance of x_2 – "explained" by the variable x_3 divided by the variance of the dependent variable. In simple regressions, the coefficient of determination is the square of the correlation coefficient.

In the case of standardized variables ($\sigma^2(x_1) = \sigma^2(x_2) = \sigma^2(x_3) = 1$), the calculations are very simple:

$\sigma^2(x_2) = 0.99^2\,\sigma^2(x_3) + \sigma^2(\varepsilon_{2.3})$,	$R^2 = 0.99^2 = 98\%$	$\sigma^2(\varepsilon_{2.3}) = 1 - R^2 = 0.02$
$\sigma^2(x_1) = 0^2\,\sigma^2(x_2) + \sigma^2(\varepsilon_{1.2})$	$R^2 = 0\%$	$\sigma^2(\varepsilon_{1.2}) = 1$
$\sigma^2(x_1) = \mathbf{0.14107}^2\,\sigma^2(x_3) + \sigma^2(\varepsilon_{1.3})$	$R^2 = \mathbf{1.99}\%$	$\sigma^2(\varepsilon_{1.3}) = \mathbf{0.981}$

On the basis of these three simple correlation coefficients and the coefficients of determination, Galton would probably have deduced, in 1886, that there was no connection between the variable x_1 and the two variables x_2 and x_3, or that this connection was negligible. Yule (1897) extended the method of linear regression to the case of several variables (referred to as multiple regression). Using the formulae of Yule (1897), we obtain the following results for the three coefficients of simple correlation in the above example. The coefficient of correlation between the variables x_1 and x_2 is denoted r_{12} (the coefficients of the multiple regression are rounded to the fourth decimal place).

$$x_1 = -7.0181\, x_2 + 7.0889\, x_3 + \varepsilon_{1.23}$$
$$R^2 = -7.0181 . r_{12} + 7.0889 . r_{13} = 7.0889 . 0.14107 = 100\%$$

$$\beta_{12.3} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} = -7.0181$$

$$\beta_{13.2} = \frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} = 7.0889$$

What a surprise! The two variables that had little or no effect on x_1 in the simple regressions now explain 100 per cent of the variance of x_1, when they are entered simultaneously into the multiple regression. The coefficients are very high for standardized variables. When the variable x_2 deviates from its mean by one standard

deviation, x_1 deviates from its mean by -7.0181 times its standard deviation, assuming x_3 to remain unchanged (the "all other things being equal" or *ceteris paribus* hypothesis). When the variable x_3 deviates from its mean by one standard deviation, x_1 deviates from its mean by 7.0889 times the standard deviation, assuming x_2 to be unchanged. These are extreme reactions of the dependent variable.

We can then calculate the parameter inflation factor (PIF). The PIF is an indicator proposed by Jean-Bernard Chatelain and Kirsten Ralf (2011). It is defined as the ratio of the parameter obtained in a multiple regression to the parameter obtained by simple regression. We can calculate the PIF for each of the explanatory variables of x_1:

$$PIF_{1,2} = \text{-}7.0181/0 \qquad PIF_{1,3} = 7.0889/0.14107 = 50.25$$

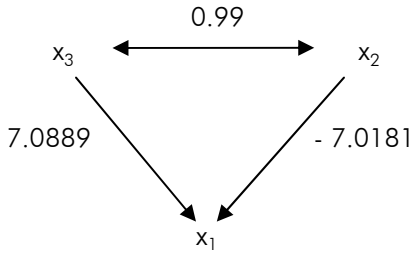
For the variable x_2, the *PIF* is infinite, while for x_3 it is 50.25. The size of the effects of the two variables x_2 and x_3 on x_1 has been amplified considerably.

Let us return to the interpretation, *ceteris paribus*, of the coefficients of the multiple regression. This was originally proposed by the US economist Henry Ludwell Moore (1917), who was, incidentally, a great admirer of Augustin Cournot's *personality* (Moore, 1905). It has since become firmly established, but the above example shows that its systematic use does not always make sense. The variables x_2 and x_3 are very highly correlated. According to the first equation, if x_2 deviates from its mean by one standard deviation, x_3 will also deviate from its mean by almost one standard deviation. Judea Pearl (2009, pp. 356–57) suggests that it is ***always*** possible to perform a counterfactual thought experiment "as if" x_3 remained unchanged, through an *intervention* of x_2, independently of the very strong statistical connection between x_2 and x_3. This also appears to be the opinion of Søren Johansen (2005) for interpreting long-term parameters in the broader context of co-integrating coefficients. But the results of these experiments must still make sense and justify very high PIFs, explaining the colossal amplifications of parameters in multiple regressions.

b. An interpretation based on causal path analysis

To better understand the problem, we shall now conduct a path analysis of the "partial" statistical connections, as proposed by the geneticist and statistician Sewall Wright (1920) (Figure 1.2).

Figure 1.2: Causal paths with mediation by x_3



x_2 has a direct partial effect on x_1, given by the coefficient -7.0181. It also has an indirect partial effect on x_1 through the mediation of x_3. This can be calculated as the product of the effect of x_2 on x_3 multiplied by the effect of x_3 on x_1, that is to say, 0.99 x 7.0889. The total effect of x_1 on x_2 is the sum of the direct and indirect effects: it is equal to the simple correlation coefficient. In the present case, the indirect effect is exactly offset by the direct effect, so that the total effect is zero (apart from the rounding errors):

$$-7.0181 + 0.99 \times 7.0889 = 0$$

There are two possible interpretations of this result.

- The multiple regression is false: the two variables x_2 an x_3 measure more or less the same thing, or are closely connected to each other. In reality, they have little or no effect on the variable x_1. The high values of the multiple regression parameters are mainly due to the strong correlation between the two explanatory variables. The result obtained is an artefact

of the multiple regression formulae obtained by Yule in 1897, when the explanatory variables are highly correlated.

- Multiple regression describes a perfectly homeostatic model, to borrow an important idea in medicine, proposed by Claude Bernard ([1865] 1993) and then developed by Norbert Wiener (1948) in his paper on cybernetics. One of the two variables, for example x_3, is a variable associated with negative feedback following a shock on the other explanatory variable, x_2. This makes it possible to achieve perfect stability of the variable x_1 despite the shock on x_2. For example, x_3 could be a variable of countercyclical monetary or fiscal policy (Hoover, 2001, pp. 45–6).

For Kevin Hoover (2001, pp. 45–6), the method of regression and the observations of variables do not allow to establish a distinction between these two "ontologically" different interpretations. Additional information about the nature of the variables is necessary to choose which interpretation should be given to the results. The spurious regressions presented here do not correspond to any of the various regressions that have been criticized since Karl Pearson (1897), as outlined in John Aldrich (1995). Herbert Simon (1954), for example, focused on spurious regressions associated with a zero partial correlation (associated with a zero parameter in a multiple regression) while the simple correlation coefficient (associated with a simple regression) is non-zero.

Our case of regression corresponds to a discordance in the opposite sense to that of Simon (1954). In our case, the partial correlation coefficient is non-zero (in fact, it is very high), while the simple correlation coefficient is zero. This situation is identical to violation of the "stability condition" of conditional independence proposed by Pearl (2009, p. 48), also called the "causal faithfulness condition" by Peter Spirtes *et al.* (2000). These authors pursued Wright's (1920) initial approach of analysing the causal paths between variables. They proposed algorithms that allow to decide whether to keep or eliminate potential causal links between variables. These algorithms are based on conditions to be satisfied (or not) by the coefficients of simple and partial correlations.

For example, the elimination of explanatory variables that have zero simple correlation with the dependent variable was the point of departure for the selection of explanatory variables in a recent version of the algorithm of Spirtes *et al.* (2000)

proposed by Peter Bühlmann *et al.* (2010). I shall go into this point in more detail in Section 4. In their article, they apply the algorithm to a sample where the number of observations is 71 bacteria producing riboflavin for 4088 explanatory variables corresponding to the same number of different gene expressions in these bacteria.

At first, the reasoning advanced by Spirtes *et al.* (2000) and Pearl (2009) is that violations of the stability condition of conditional independence are very rare. More precisely, strict equality between parameters should have a value of zero (such as: -7.01 + 0.99 x 7.08 = 0 in our example) in the whole distribution of parameters, when they are free to vary independently of each other (Pearl, 2009, p. 62). In other words, the artefact of multiple regression we are talking about here should occur very infrequently.

David Freedman (1997), on the other hand, argues that there is no reason to reject the hypothesis that the parameters are linked by equality constraints. Hoover (2001, pp. 45–6), in particular, maintains that this phenomenon is very frequent in economics, because of the possibilities of control through economic policy. He presents a theoretical diagram of the IS-LM model in which the monetary policy variable can keep economic activity unchanged following a shock from another economic factor. The question posed is not the theoretical possibility of the homeostatic model. It concerns the observation of empirical correlations where the key rates make it possible to completely and rapidly neutralize external shocks on GDP, to the point where HFP variance completely disappears. In this respect, the example is ill-chosen. The influence of central banks over economic activity is far from being reactive or effective enough to neutralize GDP variability or inflation by making them exactly non-correlated with external shocks. Wiener (1948), on the other hand, expressed his surprise on observing social and political systems that fail to self-regulate.

The present article argues that these spurious regressions occur much more frequently than Spirtes *et al.* (2000) and Pearl (2009) appear to believe. This stems from quite a different reason from the possibility of perfect control in the homeostatic models described by Hoover (2001). As I shall show, this frequency is due to (1) a disconnection between the origin of these spurious regressions and the inference test of the existence of an effect, proposed by Fisher, and (2) criteria determining the success of researchers. Many researchers publish these spurious regressions without realizing it.

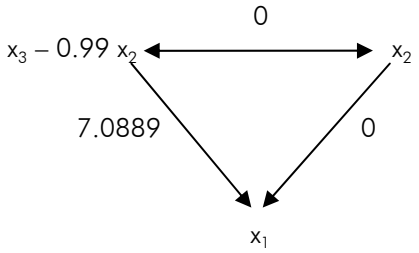
c. An interpretation by orthogonalization of the explanatory variables

To address point (1) above, I approach these spurious regressions from another angle. Since this problem is associated with excessive correlation between the explanatory variables, we can get round it by transforming the correlated variables into non-correlated variables. By analogy with Euclidean geometry, this transformation is sometimes called "orthogonalization" of the explanatory variables. By including the residuals of the regression between the two explanatory variables in the multiple regression, we obtain the following result:

$$x_1 = 0\, x_2 + 7.0889\, (x_3 - 0.99\, x_2) \qquad R^2 = 100\%$$

The new multiple regression obtained corresponds to the factoring-in of the parameter 7.08 of the first multiple regression. Unlike the first multiple regression, it clearly appears that x_2 has no effect on x_1. Only one explanatory variable remains: this regression is equivalent to a simple regression. The coefficient of the remaining variable (the variable x_3 net of its correlation with x_2) is very high (Figure 1.3):

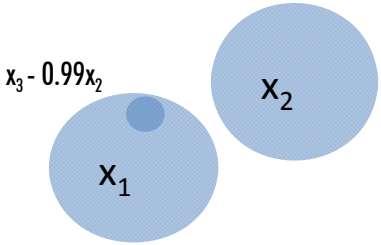
Figure 1.3: Causal paths after orthogonalization



The coefficient of determination R^2 is unchanged: it is equal to 1, as is the coefficient of simple correlation between the variable x_1 and the variable $x_3 - 0.99\, x_2$. This variable therefore "explains" the whole of the variance of x_1. The dispersion of observations of the explanatory variable ($x_3 - 0.99x_2$) around its mean is tiny. Its standard deviation is no longer equal to 1, but 0.02, in other words 2 per cent of the standard deviation of the explained variable. In the Venn diagram (Figure 1.4), the circle representing the variance of this variable is relatively small. It is entirely

included within the dependent variable x_1, because in this extreme example, the coefficient of determination R^2 is equal to 1. Lastly, the circle representing the variance of the variable x_2 has no intersection with the circle representing the variance of the dependent variable x_1 because the correlation between the two variables is zero, as in the Venn diagram before orthogonalization (Figure 1.1).

Figure 1.4: Venn diagram after orthogonalization.



The parameter of the simple regression 7.08 is now a "non-standardized" parameter. Generally, the relation between a standardized parameter (indexed by *S*) and a non-standardized parameter is the following:

$$\beta_{12} = \beta_{S,12} \frac{\sigma(x_1)}{\sigma(x_2)}$$

In the present case, the non-standardized parameter 7.08 corresponds to a standardized parameter equal to the coefficient of correlation (this is a simple regression) multiplied by the ratio of the standard deviation of the dependent variable (equal to 1) to the standard deviation of the explanatory variable.

$$7.0889 = 1 \times \frac{\sigma(x_1)}{\sigma(x_3 - 0.99x_2)} = \frac{1}{\sqrt{1-0.99^2}} = \frac{\mathrm{cov}(x_1, x_3 - 0.99x_2)}{\sigma^2(x_3 - 0.99x_2)}$$

When the dispersion of observations of the explanatory variable around its mean is relatively tiny compared with the dispersion of the dependent variable, the *size* and *sign* of the estimated parameter are very unstable, depending on whether we add or subtract an atypical observation that is far-removed from the mean of the explanatory variable. This observation can exert an upward or downward leverage effect on the value of the parameter of the regression. In this context, it is very frequent for the addition of a small number of observations to have a powerful effect on the parameter obtained, even if the R^2 is very high in the initial sample.

The interpretation of the problem has shifted with orthogonalization. It is no longer a problem of correlation between two explanatory variables. It is now a problem of an estimated parameter that is very high and very sensitive to observations with a powerful leverage effect.

This result is far from trivial. The initial introduction of two explanatory variables that are highly correlated with each other, and between which the "difference" is based on a small number of observations, makes it possible to transform a particular case into a general case. For example: "Botswana is an African country with high economic growth, unlike the other African countries. Moreover, it has received development aid" becomes "Development aid only has an effect on growth for developing countries that have 'good' macroeconomic policy". The second assertion, which is more general, is much easier to get published (Chatelain and Ralf, 2011).

Consequently, for these spurious regressions:

(1) the orthogonalization of explanatory variables brings to light the absence of connection between one of the explanatory variables and the dependent variable;

(2) the remaining explanation is associated with a residual variable (the residuals of a regression between two strongly correlated variables) of which the observations are highly concentrated around its mean. As a result, the parameter will be very high, but also very sensitive to the presence of a few observations with strong leverage effect.

Moreover, the fact that the dispersion of observations of the explanatory variable is narrow is *a priori independent* of the number of observations. This

problem of the instability of high-value parameters *must not be confused* with the question of statistical inference addressed in the following section, which takes into account the number of observations. In reality, statistical inference is of no help in solving this problem. On the contrary, the use of a large sample could delude the researcher into thinking that the high parameters are a solid result.

d. A problem neglected by the statistics and econometrics manuals?

It is important to specify the originality of the problem described in this article, compared with the way the problem of high correlation (sometimes called "multicollinearity") between the explanatory variables is presented in most modern statistics and econometrics textbooks. Implicitly, these manuals only consider cases where the simple correlations between a pair of explanatory variables and the dependent variable are not close to zero. Figure 1.5 presents the Venn diagram in this case. For Figure 1.6, I have taken an example with $r_{12} = 0.5$, and, in order to remain in the case of an exact multiple regression where $R^2 = 100\%$, $r_{13} = 0.61717$, which gives parameters of multiple regression such that $\beta_{12.3} = -5.5778$ and $\beta_{13.2} = 6.1392$. This time, the indirect effect of the explanatory variable x_2 on the dependent variable x_1, via mediation by the variable x_3, is not exactly offset by the direct effect, so that the total effect is no longer zero:

$$\beta_{12.3} + r_{13}\,\beta_{13.2} = -5.5778 + 0.99 \times 6.1392 = 0.5 = r_{12}$$

Figure 1.7 illustrates the causal paths after orthogonalization. In the usual textbook case, the orthogonalization of explanatory variables no longer leads to the conclusion that there is no connection between one of these explanatory variables and the dependent variable, unlike the artefact highlighted in this article.

Figure 1.5: Venn diagram for the three simple correlations (textbook case).

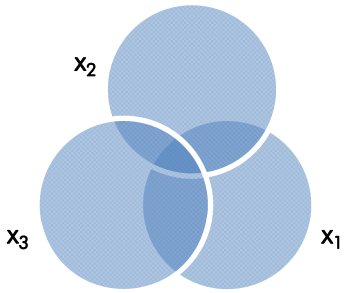


Figure 1.6: Causal paths with mediation by x_3

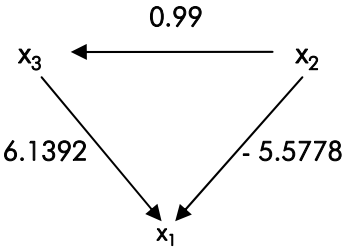
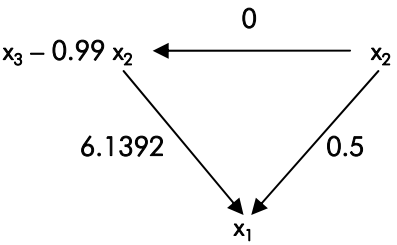


Figure 1.7: Causal paths after orthogonalization (textbook case)



2. Inference of the existence of a connection between two variables

a. From induction to inference

Induction is a method described, notably by Aristotle, for establishing general relations and predictions based on a limited number of observations (Milton, 1987). It is rejected by the sceptic philosophers, such as Sextus Empiricus ([v.200] 1997) and David Hume ([1739] 2000).

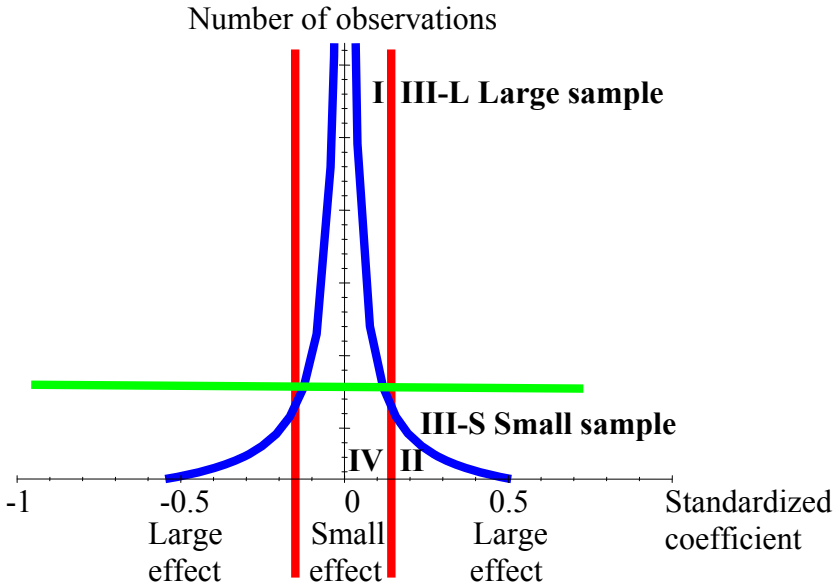
One response of statisticians and probabilists to the problem of induction is to establish inferences by associating probabilities with the frequencies of observed events (Keuzemkamp, 2000). Once the parameters connecting variables in a regression have been calculated, an inference must be made on the hypothesis of the existence of a connection between the variables. This amounts to testing the null hypothesis of the parameter associated with the two variables. The frequentist intuition is that a greater number of observations can reduce the different probabilities of being mistaken when carrying out the test.

Fisher (1925) adapted a statistics test, initially proposed by Student (the pseudonym of William Sealy Gosset) (1908), to the test of a null hypothesis of a parameter in a simple or multiple regression. Each estimated parameter of the regression is associated with an estimated standard deviation measuring the uncertainty about the value of the parameter, which is a decreasing function of the number of observations, denoted *N*. For the simple regression, the Student statistic is the ratio of the estimated parameter to its estimated standard deviation. If this statistic exceeds a pre-defined threshold, denoted $t_{N,1-\frac{\alpha}{2}}$, in practice, to the order of 1.96 when the sample is of more than 100 observations, then we can consider that there is less than one chance in 20 ($\alpha = 5\%$) of being mistaken in rejecting the null hypothesis, on the condition that, for the real model, the null hypothesis of the parameter is true (what is called the type I error probability, with the notation $p<0.05$).

$$t_{12} = \frac{r_{12}}{\sqrt{1-r_{12}^2}}\sqrt{N-2} > t_{N,1-\frac{\alpha}{2}} \qquad (2.1)$$

Compared with the previous section, a new factor comes into play: the number of observations. In Figure 2.1 below, the vertical axis represents the number of observations and the horizontal axis the correlation coefficient. The critical zone corresponding to rejection of the null hypothesis of the coefficient is above a funnel-shaped curve, given by the case of equality in equation (2.1). Within the funnel, the null hypothesis is not rejected.

Figure 2.1: Critical zone of Student's test of a simple regression



Fisher's style of inference has entered into systematic use in scientific publications in many domains of applied science: economics, epidemiology, ecology, marketing, education sciences, and so on. As suggested by the ironic title of an article by the statistician Jacob Cohen (1994) ("*The earth is round* ($p < 0.05$)"), all scientific

hypotheses must verify that the conditional probability of deciding that there is an effect – knowing that there is not (also called "type I error") – is less than 5 per cent to be considered valid.

Table 2.1 Discordance between substantive significance and statistical significance.

	Small sample (and small total population)	Large sample
Small size of the effect ("negligible" effect)	Zone IV: existence of the effect rejected and effect of negligible size	**Zone I: existence of the effect not rejected, although it is of negligible size**
Large size of the effect	**Zone II: existence of the effect rejected although it is of large size.**	Zone III: existence of the effect not rejected and effect of non-negligible size.

Such a methodological coup over the universal validation of scientific truths that establish connections between variables cannot fail to be controversial. Generally speaking, can there be one sole criterion of truth for induction or for statistical inference? The counter-argument of Sextus Empiricus ([v.200] 1997, II, 4, 19), taken up by Hume ([1739] 2000), is hard to answer: "If disagreement over a criterion is to be settled, then we must have a criterion on which we are agreed", and so on, which leads either to an infinite regression in the search for a new criterion to validate the criterion of the previous step, or to a circular argument if the first criterion is used to validate itself.

Consequently, statisticians have debated the use of criteria other than that of Fisher ever since it first appeared. Student, in the letters mentioned by Pearson (1939), and Jerzy Neyman and Egon Sharpe Pearson (1933) uphold a different opinion about the Fisher style of inference (McCloskey and Ziliak, 2008; Ziliak 2008). I shall briefly recall some of the most important critiques for statisticians.

Researchers generally wish to obtain results outside the funnel: that is where they can infer that the parameter is non-null, according to Fisher. In this case, to use Fisher's term, the parameter is said to be *statistically significant* or significantly different from zero. Fisher therefore used the word "significant", but in informal language, in both English and French, a significant effect is a strong effect. The criterion of informal language, also called "substantive" significance ("that which has meaning"), corresponds to a high correlation coefficient, independently of the number of observations in the sample, for example greater (in absolute value) than the values indicated by the two vertical lines in Figure 2.1. Fisher's statistical significance does not correspond to the same zones as substantive significance. We therefore have two types of discordance between these concepts of significance (Table 2.1).

In Figure 2.1, area II contains the cases where the effect is important but the total homogeneous population is very small. Consequently, *a homogeneous sample* is necessarily small. Orphan diseases in medicine are one example. Another is the case where one of the variables corresponds to institutional characteristics that only appear in 20 or so developed countries. Even if the effect (the correlation coefficient) is very strong, if the sample is too small, one can never infer that this parameter is statistically different from zero ($p<0.05$).

Conversely, in area I of Figure 2.1., if the sample is very large, one can end up considering that a tiny or negligible effect of the variable x_2 on x_1 is "statistically significant" ($p<0.05$). To say the effect is tiny means that a shock of one standard deviation of x_2 from its mean makes the variable x_1 deviate very little from its own mean. In this case, the statistical significance of tiny effects can be obtained by simply extending the samples indefinitely, *and aggregating very heterogeneous populations in the process*.

b. Resolving the conflict between substantive and statistical significance

One answer to this debate is to say that the simple hypothesis of the existence of an effect ($r = 0$) to be tested is not the best way to proceed. It makes more sense to test a composite hypothesis $r > r$ (*minimum threshold)*, specifying a minimum

threshold below which the effect is considered negligible. The question then is how to determine this minimum threshold. In some domains, one might consider a minimum threshold for a correlation coefficient of at least 0.1 (in other words, a coefficient of determination explaining at least 1% of the variance in a simple regression) to be reasonable. In financial activities, for example, (exchange or bond markets), very small variations in prices can lead to considerable profits. In physics, very small variations in certain phenomena can have considerable consequences for the validation of certain theories. The idea of what constitutes a negligible effect depends on the context. So what can we do?

Fisher's approach also has an arbitrary element concerning another threshold – that of the type I error probability (the p-value): why choose 5 per cent? And it neglects another p-value: the type II error probability. The type II error can be calculated when one knows the hypotheses different from the initially-tested hypothesis, (for example, all the possible values of the correlation coefficients not equal to zero). Moreover, when we vary the threshold on the type I error p-value, the type II error probability also changes.

In different letters quoted by Ethon Pearson (1939), Student proposed another criterion to be minimized as a means of deciding the threshold of the two types of p-value, and also, potentially, the minimum threshold of a parameter. He suggested a *loss function* involving two costs, each of which is specific to one of the error types (McCloskey and Ziliak, 2008; Ziliak, 2008). The introduction of these specific costs takes into account the "context" of the decision. The simplest loss function is the expected total cost of the two error types. In this case, we calculate the average cost weighted by the probabilities of each error type. This practice is, for example, similar to that used by banks when deciding whether or not to approve a loan. The cost of not giving a loan to someone who would pay it back (type II error) is the loss of the margin of intermediation multiplied by the size of the loan. The cost of giving a loan to someone who defaults (type I error) is the loss of the value of the loan plus interest. The second cost is higher than the first. The bank will therefore choose a much lower probability for type I errors than for type II errors: in other words, it will be more restrictive in its lending.

Critics of the Fisher approach argue that it does not minimize a loss function for scientific disciplines, especially in terms of the practical consequences of decisions

taken in economics, medicine, and so on. The decisions and actions that follow from results based on the Fisher criterion will not be optimal, since, by construction ("$p < 0.05$"), it does not take into account the relative costs of the two types of error.

This discussion between the Fisher criterion and a loss function associated with the consequences of a decision, as proposed by Student, is analogous to a distinction between two criteria made by Sextus Empiricus ([v. 200] 1997, 1.11.21): "Criterion has two meanings: that which we use to convince ourselves of the existence or non-existence of something (…), and that which concerns action: in attaching ourselves to it, we shall do some things and not do others". The first of these criteria is rejected by the sceptic philosophers. They do accept the second criterion, however, to avoid being condemned to inaction.

What is the connection between the spurious regression described in section 1 and the criticism of the Fisher approach presented in section 2? This regression does not correspond to either zone I or zone II. The effects obtained by multiple regression are very *strong* (when we interpret them *ceteris paribus*), and they can be *statistically significant*, even though there is no connection between x_1 and x_2. These spurious regressions are situated in zone III, which is wrongly considered to be reliable in the debate between substantive and statistical significance. So there are erroneous inferences in zone III.

There is a third phenomenon that establishes a connection between Fisher-style inference and these spurious correlations. There are other loss functions than that of the benevolent social planner, minimizing scientific errors. These are the loss functions of individual researchers whose scientific careers depend on the norms of publication criteria.

c. Stability tests of conditional independence

At the end of Section 1.c, it was mentioned that the statistical inference described in Section 2 is not suitable for addressing the statistical artefact studied in this article.

Modern statistics and econometrics textbooks emphasize the effect that very high correlation between explanatory variables has on the *estimated variance* of the estimated parameters. To this end, they propose to calculate, for example, the variance inflation factor (VIF). In the case of a regression with only two explanatory

variables, the VIF is an increasing function of the correlation coefficient between the explanatory variables, of which the growth is all the stronger as the correlation increases:

$$\mathrm{VIF}_{32} = \frac{1}{1 - r_{32}^2}$$

By doing this, the textbooks implicitly focus on a case of instability of conditional independence, where the null hypothesis $\beta_{12.3}=0$ is not rejected in the regression including an explanatory variable highly correlated with the other explanatory variables. The null hypothesis $r_{12}=0$, on the other hand, is rejected when the highly correlated variable is excluded.

Nevertheless, the artefact presented in this article corresponds to the opposite case of instability of conditional independence. The null hypothesis $\beta_{12.3}=0$ is rejected in the regression including an explanatory variable highly correlated with the other explanatory variables. The null hypothesis $r_{12}=0$, on the other hand, is not rejected when the highly correlated variable is excluded (see the examples in Tables 2.2).

In the textbooks, very high correlation becomes a problem of "statistical significance" of an effect that is rejected. It is frequently recommended that increasing the sample size can reduce the estimated variance of the estimated parameters, and therefore counteract the effect of a very high VIF. Thus, one is supposed to restore the statistical significance of the parameters despite the problem of very high correlation between explanatory variables.

To verify the interest of some of the recommendations made in the textbooks, Chatelain and Ralf (2011) conducted Monte-Carlo simulations on samples of multi-normal laws with zero mean and standard deviation of one. The variable x_2 has a theoretical zero correlation r_{12} with the variable x_1. The computer simulations of random samples, the numerical values of which are calculated to the nearest twelfth decimal place, are such that we never obtained a correlation exactly equal to zero. For each sample, however, one can perform inference tests on the hypotheses $r_{12}=0$ and $\beta_{12.3}=0$. One can then calculate the percentages of stability or instability of the conditional independence. These are presented in Tables 2.2.

It can be observed that when the correlation between the explanatory variables is high, but not too high ($r_{23}=0.50$), the tests conclude that the conditional

independence associated with the null hypothesis of the effect of variable x_2 is stable for more than 90 per cent of the samples. Moreover, this percentage varies little with the increase in the number of observations of each sample (from 92.3 per cent for 102 observations to 90.6 per cent for 1002 observations).

Tables 2.2. Inferences on the stability of conditional independence for 1000 random samples of multi-normal laws

Table 2.2.1: $r_{23} = 0.50$, N=102 observations

	Does not reject $r_{12} = 0$	Rejects $r_{12} = 0$
Does not reject $\beta_{12.3} = 0$	92.3%	3.3%
Rejects $\beta_{12.3} = 0$	2.1% ("artefact")	2.3%

Table 2.2.2: $r_{23} = 0.50$, N=1002 observations

	Does not reject $r_{12} = 0$	Rejects $r_{12} = 0$
Does not reject $\beta_{12.3} = 0$	90.6%	2.4%
Rejects $\beta_{12.3} = 0$	4.9% ("artefact")	2.1%

Table 2.2.3: $r_{23} = 0.99$, N=102 observations

	Does not reject $r_{12} = 0$	Rejects $r_{12} = 0$
Does not reject $\beta_{12.3} = 0$	42.3%	2.8% ("textbooks")
Rejects $\beta_{12.3} = 0$	52.3% ("artefact")	2.8%

Table 2.2.4: $r_{23} = 0.99$, N=1002 observations

	Does not reject $r_{12} = 0$	Rejects $r_{12} = 0$
Does not reject $\beta_{12.3} = 0$	0%	0% ("textbooks")
Rejects $\beta_{12.3} = 0$	95.5% ("artefact")	4.5%

Note: The random samples of multi-normal laws have a very low correlation with the explanatory variables $r_{12} = 0$, $r_{13} = -0.03$, varying the correlation between the explanatory variables r_{23} and the number N of sample observations (Chatelain and Ralf, 2011).

On the other hand, when the correlation between explanatory variables is very high ($r_{23} = 0.99$), the tests conclude that the conditional independence associated with the null hypothesis of the effect of variable x_2 is stable for only 45.3 per cent of the samples of 102 observations and for only... 0 per cent of the samples of 1002 observations. This is associated with 52.5 per cent of cases of artefact for the samples of 102 observations, and 95.5 per cent of cases of artefact for the samples of 1002 observations.

Consequently, recommending that the number of observations should be increased actually favours the inference of artefacts when the correlation between explanatory variables is very high and when the correlations between the explanatory variables and the dependent variable are low.

When the textbooks emphasize the effect of increasing the estimated standard deviations of estimated parameters, they rarely associate this with a warning about the fragility of the *ceteris paribus* interpretation of the size of these statistically significant but very high parameters.

This marks a regression in the quality of statistical practices compared with the 1940s and 1950s. During that period, the "confluence analysis" developed by Ragnar Frisch (1934) was regularly used. Among other things, it indicated problems of high correlation between explanatory variables when a bunch map "exploded" (Tinbergen, 1939; Hendry and Morgan, 1989; Armatte, 2001). Jan Tinbergen (1939, pp. 28–31) proposed the following practice. As a first step, one could carry out a test of "statistical significance" of the estimated parameters, following the Fisher approach. But, *in a second step,* in the event of a high correlation between variables being given by an "exploded" bunch map, the estimated parameters that are too high should then be considered "uncertain" and not accepted, *even if these estimated parameters are "statistically significant"*.

In this case, one can either remove an explanatory variable that is highly correlated with the others and considered a "nuisance parameter", according to Tinbergen (1939), or keep it by adding a constraint on the values of the parameters of the highly correlated variables. If these practices have regressed, it is perhaps because other forces are at work, pressuring researchers to obtain "statistically significant" effects.

3. Artefact of publication, meta-analysis and spurious regressions

a. Artefact of publication and meta-analysis

In this section, I describe the consequences of the Fisher criterion for researchers. Fisher-style inference became established in many scientific disciplines during the 1950s and 1960s. When the Student test gives a result that rejects the null hypothesis of partial correlation, a norm imposes itself among the editors of scientific journals. The only results to be published are those that reject the null hypothesis of the parameter measuring the effect between two variables. This creates a statistical artefact in the selection process of articles published, often referred to as "publication bias". This is an extension of the use of the word "bias" by statisticians to denote the gap between an estimated parameter and the "true" value of the parameter, in various situations. Results that do not reject the null hypothesis of the parameter are sometimes called "negative" results (of the test), even if the sign of the parameter is positive.

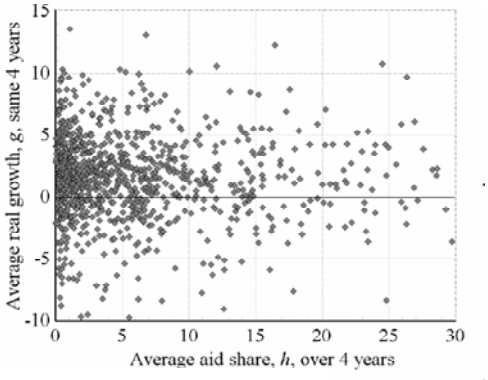
In the 1950s, in medicine, researchers decided to conduct "meta-analyses", compiling the estimated parameters of a relation between two particular variables in all the studies available – studies using different samples and different estimation methods, carried out by researchers whose *a priori* are not necessarily the same, and so on. They established weighted averages of these estimated effects and their statistical dispersion.

In a meta-analysis, one can also detect and correct the artefact of publication, because one has two dimensions: the size of the sample and the size of the effect for different studies. If, on average, studies on small samples have higher parameters than studies on large samples, one deduces that the statistical method used by the researchers in small-sample studies has presented an artefact tending to make the parameter statistically significant (Stanley, 2005). In Figure 2.1, one can identify an artefact of selection if the area III-L – corresponding to strong effects for large samples – is relatively empty compared with the area III-S – corresponding to strong effects for small samples, and the area I – corresponding to small effects for large samples. These three areas share the critical zone of rejection of the null hypothesis.

Statistical theory states that we should not obtain a relation between the estimated parameters and the number of observations. On the other hand, the estimated standard deviation of the estimated parameters decreases with the number of observations. In the presence of a confirmed artefact of publication, a calculation of the average effect from the results of different studies should underweight the parameters of small-sample studies relative to the parameters of large-sample studies.

As an example, I present the results of a meta-analysis of the statistical relation between development aid and economic growth (Doucouliagos and Paldam, 2009). Figure 3.1 presents all the observations of these two variables for a large number of countries. The shape of the scatter plot indicates that a linear regression line passing roughly through the middle of the cloud of points would be horizontal. The coefficient of simple correlation between the two variables is zero. The studies tend to focus attention on multiple regressions where the parameters could be non-null.

Figure 3.1: Simple correlation between development aid and growth

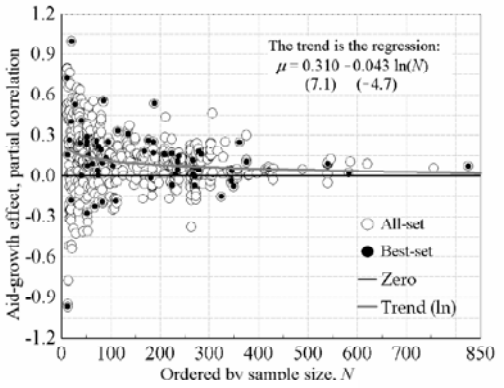


Source: Doucouliagos and Paldam, 2009, p.438.

Figure 3.2 represents the estimated parameters in different studies as a function of the sample size of each study. It can be seen that as the sample sizes increase, the dispersion of the estimated parameters decreases, following the shape of a funnel. This is an expected result: the greater the number of observations, the smaller the estimated standard deviation of the estimated parameters. On the other

hand, there is a decreasing relation between the size of the estimated parameter and the number of observations. This phenomenon should not appear: the estimated parameters used in Section 1 have no relation to the number of observations in the formulae given by Yule (1897). This unexpected result suggests that there is an artefact of publication: small-sample studies have higher parameters.

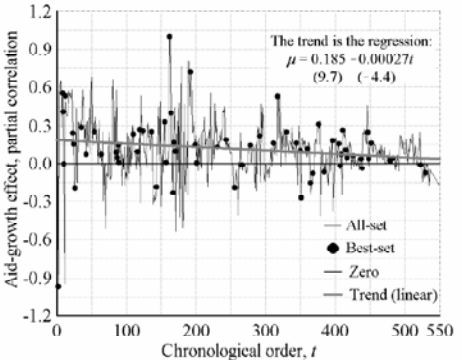
Figure 3.2: Artefact of publication: the size of the effect decreases with the size of the sample



Source: Doucouliagos and Paldam, 2009, p.452.

Figure 3.3 shows that there is also a temporal artefact of publication: the size of the published effect tends to decrease as a function of the chronological order of publication. Thus, we have what John Ioannidis (2008) associates with a result called the "winner's curse" in auctions. Researchers "overbid" the size of *new and unexpected* effects in order to get published in prestigious scientific journals. In doing so, they run a considerable risk that subsequent studies reproducing their work in less prestigious journals will gradually show that the effect is absent. Ultimately, this result will become obvious when the first meta-analyses are published several years later, when at least 30 publications are available. This process generates controversy, increasing the number of citations of the original article and the visibility of its authors. Through a positive feedback effect, this process confirms *a posteriori*, through the number of citations, the quality of the journal, in turn reinforcing the phenomena of overbidding of new, unexpected (or even bizarre), strong effects in subsequent publications.

Figure 3.3: Temporal artefact of publication: the size of the effect decreases over time



Source: Doucouliagos and Paldam, 2009, p.452.

b. Spurious regressions, appreciated regressions

It is at this stage that our spurious regressions become appreciated by researchers seeking greater visibility. They have four advantages.

(1) The parameters are high.

(2) They allow to obtain new and unexpected effects among the set of effects that the community of researchers considers *a priori* as null (rightly so, since they are truly null, except in the case of the homeostatic model).

(3) The size and sign of the estimated parameter are very sensitive to the addition or removal of a few atypical observations with a strong leverage effect (far-removed from the mean of observations of the explanatory variable). These instabilities on the *size* and (even better) the *sign* of the effect fuel the controversy, increasing the scientific visibility of the authors and the journals that publish them over the next 15 years or so.

(4) *Above all, they can be "statistically significant", and therefore publishable.* All the statistics textbooks consider that the problem of high correlation between explanatory variables is no longer a problem when the estimated parameters are statistically significant. As a solution, they frequently suggest increasing the sample

size. The problem is then reduced to the role of the inflation of the estimated variance of an estimated parameter, measured by the variance inflation factor (VIF).

On the other hand, the effect of the strong correlation between explanatory variables on the estimated parameter, measured by the "parameter inflation factor" (PIF) (Chatelain and Ralf, 2011) is neglected. *Our central argument is that obtaining the statistical significance of Fisher for the parameters of the multiple regression is no guarantee against the spurious nature of multiple regressions described in Section 1.*

How can we build a false regression, *when the simple correlation of x_2 on x_1 is close to zero*, as in the case of development aid and economic growth (see Figure 3.1)? Table 3.1 proposes four ideas for finding another explanatory variable x_3, highly correlated with x_2, in order to obtain a false regression.

Table 3.1. Construction of highly correlated pairs of explanatory variables.

x_3 highly correlated with x_2	Interest of the model:
Indicators measure two phenomena that are similar or share the same cause.	Example: opinion of experts on corruption and then on the risk of expropriation in a given country. Search for precision in the differentiation of effects.
Delayed term: $x_3 = x_2(t-1)$	Dynamic model: distinguish between short-term effects, delayed effects and long-term effects.
Powers: $x_3 = (x_2)^2$ $x_3 = (x_2)^3$	Non-linear model with increasing or decreasing marginal effects. Second, third or fourth order polynomial approximation of any given non-linear relations.
Term of interaction $x_3 = x_2 * x_4$	Complementarity, modelling of interdependence escaping from the "ceteris paribus" hypothesis.

The first example corresponds to the case where there are several explanatory variables measuring phenomena associated with very similar phenomena, or when a large number of explanatory variables are available: in this case, we are sure to find two variables that are highly correlated with each other.

The second example makes it possible to obtain a strong and, for example positive, short-term effect, offset by a strong effect of the same variable measured at the previous period and of the opposite, elevated sign. The long-term effect of aggregating these two effects will be close to zero. In this case, we will have brought to light a fallacious short-term effect.

The estimation of a non-linear model is interesting, because it is possible to measure increasing and/or decreasing marginal effects. Unfortunately, the further away from zero the mean of the observations is, the higher the correlation between a variable and the same variable squared (or to any other power). When the simple correlation between x_2 and x_1 is close to zero, one can obtain an unfounded non-linear model, the parameters of which are statistically significant.

The estimation of a model with an interaction term is very rich in lessons, because it takes into account the possible complementarity between two explanatory variables, and by construction rejects the possibility of an intervention *ceteris paribus* of one of the two variables. By construction, the interaction term is often highly correlated with at least one of the two variables. When the simple correlation between x_2 and x_1 is close to zero, one can easily obtain a model with an unfounded interaction term, the parameters of which are statistically significant. All that is required is then to invent an interesting narrative around that interaction.

4. "Pifometry" to the aid of econometrics

a. The PIF and tests on the coefficients of simple correlation

What can be done to detect the results of spurious regressions? Ioannidis (2008) proposes calculating what he calls a "vibration ratio", in other words the ratio of the size of effects in different studies or in different statistical tables of the same article, divided by the smallest estimated effect. A high vibration ratio is a signal of instability of the estimated effect. Ioannidis (2008) also specifies that this volatility

and instability of the size of estimated effects will be frequent in domains where investigation is still in its early stages and the way to study the phenomenon has not been clearly defined.

Chatelain and Ralf (2011) suggest using the Parameter Inflation Factor (PIF). This is the ratio of the parameter obtained by multiple regression divided by the parameter obtained by simple regression. If the PIF is greater than 2, then the parameter of multiple regression is more than twice the parameter of simple regression. Two calculations have already been presented in the example in Section 1.

Unlike the VIF, which only brings into play the coefficients of simple correlation between explanatory variables, the PIF also involves the coefficients of simple correlation of the explanatory variables with the dependent variable. Using the notation of Yule (1897), we denote r_{12} the coefficient of simple correlation between the variables x_1 and x_2. For a multiple regression where the dependent variable x_1 is explained by the two variables x_2 and x_3:

$$VIF_{32} = \frac{1}{1 - r_{32}^2}$$

and

$$PIF_{12} = \frac{1}{r_{12}} \frac{r_{12} - r_{13} r_{32}}{1 - r_{32}^2} = \left(1 - r_{32} \frac{r_{13}}{r_{12}} \right) \times VIF_{32} \qquad (4.1)$$

The VIF contributes to the inflation of the parameter measured by the PIF by amplifying the interval $r_{12} - r_{13}\, r_{32}$. In the calculation of the parameter, this interval corresponds to the contribution of x_2 in explaining the variance of x_1 net of the indirect effect of x_2 on x_1 via the mediation of the other variable x_3.

The PIF is a tool for editors, referees and readers of articles using regression in scientific journals. To calculate the PIF, it is necessary for the authors of the article to present, in addition to the results of their multiple regression, the number of observations, the means, the standard deviations, and the matrix of simple correlation of their variables. A large proportion of articles fail to present the matrix of correlations, despite the initial recommendation of Yule (1897).

For the authors of scientific articles, we propose to test the null hypothesis of the coefficient of simple correlation between the dependent variable and each of the explanatory variables. It is also possible to test a composite hypothesis, such that the coefficient of simple correlation should not be smaller than 0.1, for example (Chatelain and Ralf, 2011). If the coefficient turns out to be too small, we decide not to take into account that explanatory variable in the rest of the study. This preliminary condition is included in certain versions of the algorithm that can be used to draw causal graphs following the method of Spirtes *et al.* (2000).

b. Application: development aid, macroeconomic policy and economic growth

I apply these two tools to an article published by Craig Burnside and David Dollar (2000) in the *American Economic Review*. Over 10 years, this article has become one of the most cited of the articles published in that journal in 2000. To give some idea of the impact of this article, at the beginning of June 2011, there were more than 2390 citations of the article referenced in the database of articles and academic works used by Google Scholar. In the article, the authors show that development aid can only have a positive effect on growth in the presence of good macroeconomic policies. What do they mean by that? Low inflation, low budget deficit and wide opening to international trade. More precisely, the "macroeconomic policy" variable is defined by:

$$Policy = 1.28 + 6.85\ government\ budget\ surplus - 1.40\ inflation\ rate + 2.16\ (exports + imports/GDP)$$

The policy implication is the following: if the purpose of development aid is to increase economic growth, then it should only be given to developing countries pursuing "good macroeconomic policies". The authors' results are presented in Table 4.1.

The explained variable is economic growth, and the table presents three explanatory variables. The numbers between brackets below the estimated parameters are the estimated standard deviations of the parameters. If the ratio of these two values exceeds 1.96, according to the Fisher test, then there exists an effect

with a type 1 error probability of less than 5 per cent ($p<0.05$). It is customary to add an asterisk when a parameter is "*statistically significant*". In the first column, if development aid appears on its own, there is no statistically significant effect on growth, as expected in Figure 3.1. In the second column, the authors use one of the tools from Table 3.1: the addition of a term of interaction between aid and the indicator of macroeconomic policy. They still do not obtain *"statistically significant"* coefficients. In the third column, they use another tool from Table 3.1: the addition of a squared term for aid in interaction with the indicator of macroeconomic policy. This time the authors obtain two statistically significant parameters.

Table 4.1. Effect of development aid and macroeconomic policies on economic growth

Aid/GDP	0.034 (0.12)	0.015 (0.012)	0.049 (0.12)
(Aid/GDP) . Policy	-	0.013 (0.049)	0.20* (0.09)
$(Aid/GDP)^2$.Policy	-	-	-0.019* (0.0084)

Source: Burnside and Dollar, 2000
Note: for $N=365$ observations.

To calculate the indicators that I have just proposed, we start by calculating the coefficients of the simple regression based on data that can be downloaded from the internet. I use the index 1 for the dependent variable (economic growth), the index 2 for the variable *(Aid/GDP).Policy*, and the index 3 for the variable *$(Aid/GDP)^2$.Policy*. The results are the following:

$PIF_{12}=0.20/0.095 = 2.13$
$PIF_{13}=-0.019/0.0046 = -4.15$ (with change in the sign of the effect).
$r_{12}=0.13$: the hypothesis $r_{12}=0$ is not rejected ($p<0.05$).
$r_{13}=0.06$: the hypothesis $r_{13}=0$ is not rejected ($p<0.05$).
$r_{23}=0.92$.

I can calculate Ioannidis's vibration ratio for aid/GDP in interaction with policy by taking the parameters of the second line of Table 4.1: 0.20/0.013 = 15.4. All these indicators confirm that this is a spurious regression of the type described in Section 1.

In the example in Section 1, it was clear that the parameters of the pair of highly correlated variables were offsetting each other (7.08 and -7.01) because the two variables had been standardized (they had the same standard deviation equal to 1). This offsetting cannot be detected in Table 4.1. This is because authors of articles using regression usually present parameters for non-standardized variables:

$$\beta_{12} = \beta_{12}^{S} \frac{\sigma(x_1)}{\sigma(x_2)} = 0.20 \text{ and } \beta_{13} = \beta_{13}^{S} \frac{\sigma(x_1)}{\sigma(x_3)} = -0.019$$

Because of the squared term for aid, the standard deviation of the variable indexed 3 is greater than that of the variable indexed 2. Consequently, its non-standardized parameter could be much smaller than that of the variable indexed 2.

The presentation of standardized coefficients, which is provided for in most statistics programmes, is therefore another tool that can signal the problem of false correlation. In the case of simple regression, the standardized coefficient is equal to the coefficient of correlation, which lies somewhere between -1 and 1. In multiple regression, when a standardized coefficient exceeds 1 in absolute value, one can consider that the size of this parameter has been inflated because of highly correlated explanatory variables.

The article by Burnside and Dollar (2000) is an exemplary case of an article suffering from the "winner's curse". It highlights a very strong and very fragile effect in a particularly sensitive domain of economic policy: development aid. Shortly after its publication, it was the subject of controversy when William Easterly *et. al.* (2004) did not find the effect obtained by Burnside and Dollar (2000) after adding about 80 observations. The article then served as a reference for a large number of articles seeking to obtain a conditional effect of development aid on growth by introducing other explanatory variables highly correlated with aid, along the lines of Table 3.1. Finally, 15 years after the working paper was first circulated in 1995, a meta-analysis by Hristos Doucouliagos and Martin Paldam (2010) confirmed the absence of an

effect of aid conditional on the variables of economic policy among the studies published after their article.

c. Factors determining the returns on financial assets

In finance and accounting, there is an abundance of literature seeking to evaluate the factors determining the profitability of stocks and bonds. These factors include, for example, the earnings yield of a stock index for the market on which the company is listed, company size, share price to profit ratio, accounting ratios of profitability and indebtedness, past returns on the shares, macroeconomic or sector growth rates, interest rates, and so on), and lastly indicators reflecting anomalies on capital markets (for example, January always sees share prices drop in the USA).

This literature uses a flood of stock market and accounting data, accumulated notably by the Center for Research in Security Prices (CRSP) in Chicago, which celebrated its fiftieth anniversary in 2010. It often happens that factors contribute only weakly to growth of the coefficient of determination R2 (and therefore to explaining the variance in share profitability), but that they are nevertheless statistically significant because the number of observations of the sample is large (see zone I of Figure 2.1, and Table 2.1). Sometimes, the estimated standard deviations of the estimated parameters are not calculated appropriately (Petersen, 2009), which detracts from the reliability of the inferences.

More precisely, Seung Ahn, Christopher Gadarowski and M. Fabricio Perez (2009) show that two very widely cited studies – Fama and French, 1993; Jagannathan and Wang, 1996 – with more than 5900 and 1290 citations respectively on Google Scholar in June 2011, use explanatory variables (in this case, "beta" parameters estimated in a first step) with high correlation between themselves and low cross-sectional variability (the second step in these articles is a cross-sectional regression). This last property suggests weak and/or unstable simple correlations with the dependent variable, according to whether observations with strong leverage are added or removed. We are then in a case similar to the artefact described in the present article. By carrying out simulations, Ahn, Gadarowski and Perez (2009) find that the parameters evaluating the risk of these factors can be biased by 60 per cent compared with their true value.

Conclusion

It is easy to solve the problem of unfounded correlations presented in this article. I recommend that explanatory variables should not be taken into account if their correlation coefficients with the dependent variable are too small; all the more so when they are highly correlated with each other.

These spurious regressions are not obtained intentionally, for the arguments I have advanced in this article are unknown. They emerge as the result of an evolutionary process of trial and error, with the aim of obtaining statistically significant – and therefore publishable – parameters.

This succession of trials and errors is another major problem for validating the results of inference tests drawn from these regressions. This is also known as the problem of *multiple comparisons*. The researcher systematically tests a large number of explanatory variables until he obtains a statistically significant result. He follows the method of Professor Shadoko and the plumber for launching the Shadok rocket into space: "It is only by continual effort that one eventually succeeds". Successive multiple comparisons increase the type I error probability (Denton, 1985). By acting as if there had been no prior sequence of multiple comparisons, an inference using a threshold of 5 per cent for the type I error probability is therefore erroneous. This problem also leads to unfounded inferences, which are not always the same as the spurious regressions examined in this article.

References

Aldrich, J. (1995), "Correlations Genuine and Spurious in Pearson and Yule", *Statistical Science*, 10(4), pp. 364–76.

Ahn, S.C., C. Gadarowski and M.F. Perez, 2009, "Effects of Beta Distribution and Persistent Factors on the Two Pass Cross-Sectional Regression", Working paper, Arizona State University.

Armatte, M. (2001), "Le statut changeant de la corrélation en économétrie (1910-1940)", *Revue Economique*, 52(3), pp. 617–31.

Bernard, C. [1865] (1993), *Introduction à l'étude de la médecine expérimentale*, Paris: Champs, Flammarion.

Burnside C. and D. Dollar (2000), "Aid, Policies and Growth", *American Economic Review*, 90(4), pp. 847–68.

Bühlmann P., M. Kalisch and M.H. Maathuis (2010), "Variable Selection in High-Dimensional Models: Partially Faithful Distributions and the PC-Simple Algorithm" *Biometrika*, 97, pp. 261–78.

Chatelain, J.B. and K. Ralf (2011), "Spurious Regressions and Near-Multicollinearity, with an Application to Aid, Policies and Growth", working paper.

Cohen, J. (1994), "The earth is round ($p < .05$)", *American Psychologist,* 49(12), pp.997–1003.

Denton, F.T. (1985), "Data Mining as an Industry", *The Review of Economics and Statistics*, 67(1), pp. 124–27.

Doucouliagos, H. and M. Paldam (2009), "The Aid Effectiveness Literature: The Sad Results of 40 Years of Research", *Journal of Economic Surveys*, 23(3), pp. 433–61.

Doucouliagos, H. and M. Paldam (2010), "Conditional Aid Effectiveness: A Meta Study", *Journal of International Development,* 22(4), pp. 391–410.

Easterly, W., R. Levine and D. Roodman (2004), "New Data, New Doubts: A Comment on Burnside and Dollar's 'Aid, Policies, and Growth' (2000)", *American Economic Review* 94(3), pp. 774–80.

Fama, E.F., and K.R. French (1993), "Common Risk Factors in the Returns on Stocks and Bonds", *Journal of Financial Economics,* 33, pp. 3–56.

Fisher, R. (1925), "Applications of 'Student's' distribution", *Metron*, 5(3), pp. 90–104.

Frisch, R. (1934), "Statistical Confluence Analysis by Means of Complete Regressions Systems", Publication no 5, University Institute of Economics, Oslo.

Freedman D. (1997), "From Association to Causation via Regression" in V.R. McKim and S.P. Turner (eds), *Causality in Crisis?,* Notre Dame, IN: University of Notre Dame Press, pp. 113–61.

Galton, F. (1886), "Regression Towards Mediocrity in Hereditary Stature", *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, pp. 246–63.

Hendry, D.F. and M.S. Morgan (1989), "A Re-Analysis of Confluence Analysis", *Oxford Economic Papers,* 41, pp. 35–52.

Hoover, K.D. (2001), *Causality in Macroeconomics*, Cambridge, UK: Cambridge University Press.

Hume, D. [1739] (2000), *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, republished in D.F. Norton and M.J. Norton (eds), *A Treatise of Human Nature*, Oxford: Oxford University Press.

Ioannidis, J.P.A. (2008), "Why Most Discovered True Associations Are Inflated", *Epidemiology*, 19(5), pp. 640–48.

Jagannathan, R. and Z. Wang (1996), "The Conditional CAPM and the Cross-Section of Expected Return", *Journal of Finance*, 51, pp. 3–53.

Johansen S. (2005), "Interpretation of Cointegrating Coefficients in the Cointegrated Vector Autoregressive Model", *Oxford Bulletin of Economics and Statistics*, 67(1), pp. 93–104.

Keuzenkamp H. (2000), *Probability, Econometrics and Truth: The Methodology of Econometrics.* Cambridge: Cambridge University Press.

Legendre, A.M. (1805), *Nouvelles méthodes pour la détermination des orbites des comètes*, Paris: Courcier.

Magnus J.R. and M.S. Morgan (1999), *Methodology and Tacit Knowledge: Two Experiments in Econometrics*, Chichester: John Wiley & Sons Ltd.

McCloskey, D. and S.T. Ziliak (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, Ann Arbor, MI: University of Michigan Press.

Milton J.R. (1987), "Induction before Hume", *The British Journal for the Philosophy of Science*, 38 (3), pp. 49–74.

Moore, H.L. (1905), "The Personality of Antoine Augustin Cournot", *The Quarterly Journal of Economics*, 19 (3), pp. 370–399.

Moore, H.L. (1917), *Forecasting the Yield and the Price of Cotton*, New York: Macmillan.

Neyman, J. and E.S. Pearson (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses", *Philosophical Transactions of the Royal Society London (A),* 231, pp. 289–337.

Pearl, J. (2009), *Causality: Models, Reasoning and Inference* (2nd edition), Cambridge: Cambridge University Press.

Pearson, K. (1897), "On a Form of Spurious Correlation that May Arise when Indices Are Used in the Measurement of Organs", *Proceedings of the Royal Society London Series.* A, 60, pp. 489–98.

Pearson, E.S. (1939), "'Student' as Statistician", *Biometrika,* 30(3/4), pp. 210–50.

Petersen, M.A. (2009), "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches", *Review of Financial Studies,* 22(1), pp. 435–80.

Sextus Empiricus [v. 200] (1997), [Πυῤῥώνειοι ὑποτύπωσεις], *Esquisses pyrrhoniennes*, translated by P. Pellegrin, Paris: Le Seuil.

Simon, H. (1954), "Spurious Correlation: A Causal Interpretation", *Journal of the American Statistical Association*, 49, pp. 467–92.

Spirtes, P., C.N. Glymour and R. Scheines (2000), *Causation, Prediction, and Search*, 2nd edition, Cambridge, UK: Cambridge University Press.

Stanley, T.D. (2005), "Beyond Publication Bias", *Journal of Economic Surveys*, 19, pp. 309–45.

Student (1908), "The Probable Error of a Mean", *Biometrika*, 6, pp. 1–25.

Tinbergen, J. (1939), *Statistical Testing of Business Cycle Theories: A Method and Its Application to Investment Activity,* 1, Geneva: League of Nations.

Tobin, J. (1950), "A Statistical Demand Function for Food in the USA", *Journal of the Royal Statistical Society,* Series A, 113, Part II, pp. 113–49.

Tobin, J. (1999), "My 1950 Food Demand Study in Retrospect", in Magnus J.R. and M.S. Morgan (eds), *Methodology and Tacit Knowledge*, Chichester: John Wiley & Sons Ltd, pp. 265–68.

Wiener, N. (1948), *Cybernetics, or Control and Communication in the Animal and the Machine*, Cambridge, MA: MIT Press.

Wright, S. (1920), "The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs", *Proceedings of the National Academy of Sciences*, 6, pp. 320–32.

Yule, G.U. (1897), "On the Theory of Correlation", *Journal of the Royal Statistical Society*, 60, pp. 812–54.

Ziliak, S.T. (2008), "Guinessometrics: The Economic Foundation of 'Student's' t", *Journal of Economic Perspectives*, 22(4), pp. 199–216.

Reply by Xavier Ragot (CNRS)

I find this article fascinating in its approach. It identifies a statistical problem exacerbated by a sociological dynamic of the profession, resulting in the erroneous production of scientific knowledge. There is a global artefact in the discipline, especially in mine – economics: the statistical problem is double and perfectly identified, and the social scientists of the *milieu* make it worse.

In addition, Jean-Bernard has the solution, in the shape of the PIF, to get this research back on the right track. It is therefore an article with Promethean ambition, and I would like to see it applied on a large scale: to use the PIF on the meta-analyses already conducted to see how effective it is as a statistical method of identification. This is the extremely inspiring nature of the article. There is a vast field open to it.

As for the way it is written, it is certainly dense: it contains some history of econometrics, sociology of science, statistics....

The depressing side is that necessarily, in this discipline, especially in econometrics, we know *a priori* that we are going to identify small effects with explanatory variables that will be highly correlated. We are therefore, structurally, almost in the case that Jean-Bernard describes. When we take a developing country, everything is correlated: there is no level of education, no infrastructure; there are diseases. Everything is correlated in the explanatory variables. We are nevertheless going to try to identify development aid, because we have an *a priori* that it is useful, and so, of course, and unfortunately, you tell us that even if there is an effect, scientifically, we cannot say so.

Finally, I therefore deduce that aid development to poor countries cannot be justified by science, given our current sample size. We either give it or we don't, but the data do not allow us to settle the question one way or the other.

With regard to econometrics, you open up a vast field of undecidability of correlations – I won't even speak of causalities. There is an endeavour currently under way, if we take the latest Clark medal winner, to examine all public policy through the filter of econometric evaluation: one should only do what has been evaluated. This article, on the other hand, says that that is not a promising programme of research, because the field of undecidability of our statistical methods is much too large, and so even if we correct the PIFs, we cannot determine causalities.

Ultimately, we are condemned to the statistical poverty of causality, which is the pessimistic conclusion of the statistician.